Integrative RNA modeling
Structure Probing

Ivo Hofacker

Institute for Theoretical Chemistry, University of Vienna

http://www.tbi.univie.ac.at/

AlgoSB 2025
Marseille, December 2025

Experimental Structure Determination

High resolution structure determination

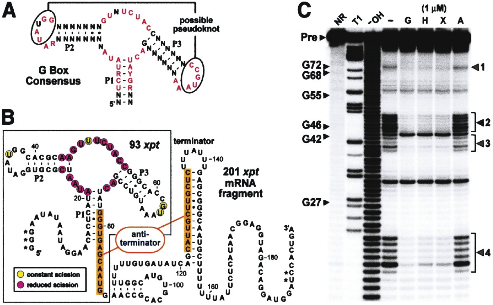
- X-ray Crystallography – requires crystals
- NMR spectroscopy – small RNAs only
- Cryo-EM – best for large RNA-protein complexes
- FRET – distance between just two fluorescent probes

Cheap alternative:
Structure probing:

- RNA is chemically modified in a structure dependent manner
- Readout via sequencing modified RNA

In-line probing

- Chain breaks happen spontaneously at room temperature
- Break points mostly in unpaired regions
- → structure dependent cleavage
- Can be accelerated by adding lead

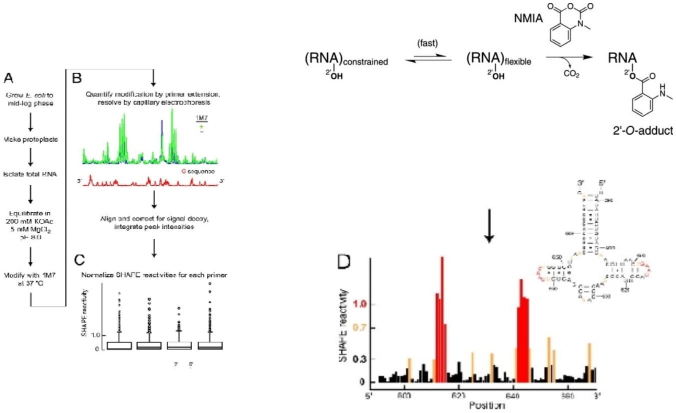


[1]

[1] Mandal, . . ., Breaker, 2003

SHAPE Probing

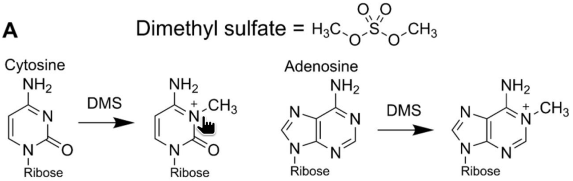
Selective 2’-hydroxyl acylation analyzed by primer extension


NMIA
(fast)
(RNA)constrained
2' OH
(RNA)flexible
2' OH
CO2
RNA
2'-O-adduct
A
Grow E. coli to mid-log phase
Make protoplasts
Isolate total RNA
Equilibrate in 200 mM KOAc 5 mM MgCl2 pH 8.0
Modify with 1M7 at 37 °C
B
Quantify modification by primer extension, resolve by capillary electrophoresis
1M7
+
−
G sequence
5'
3'
Align and correct for signal decay, integrate peak intensities
C
Normalize SHAPE reactivities for each primer
SHAPE reactivity
1.0
0
D
SHAPE reactivity
1.0
0.7
0.3
0
5'
600
620
640
660
3'
Position
3'
5'

- Probes flexibility rather than base pairing
  paired nucleotides → C3’-endo sugar pucker → low reactivity
- Little sequence bias
- Several different reagents available (1M7, NMIA, NAI)

Deigan, Li, Mathews, Weeks 2009

DMS (Dimethyl Sulfate) Probing



- Modification on the WC edge
- Directly probes base pairing
- Mostly probes A and C, no data for G and U

Enzymatic Methods

- Use enzymes that cleave only single strand / double stranded regions
  often a pair of single strand / double strand specific enzymes
- Typically probes only specific nucleotides
- Only sites accessible to a bulky protein
- Not usable *in vivo* (in contrast to SHAPE and DMS)

Not as widely used anymore

How to measure Reactivity

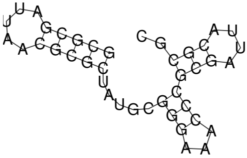
- Old-school: Gel-electrophoresis
- Modern: Readout via sequencing
  - SHAPE-Seq: Modification causes RT-stop
  - SHAPE-MaP: Modification causes mis-incorporation (mutation)

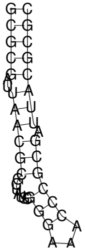
Perform experiment with and without reagent
measure mutation (or RT-stop) frequency m_i at pos i

$$\text{reactivity: } r_i = \frac{m_i^{\text{treated}} - m_i^{\text{untreated}}}{m_i^{\text{denatured}}}$$

MaP (mutational profiling) allows multiple mutations in a single read!

Incompleteness





```
GCGCGAUUAACGCGCUAUGCGGGAAACCCGCGAUUACGCGC
(((((.....)))))...(((((...)))(((....)))))    -9.30
(((((.....(((((...))(((...))))))....)))))    -8.50
XXXXX.....XXXXX...XXXXX...XXXXXX....XXXXX
```

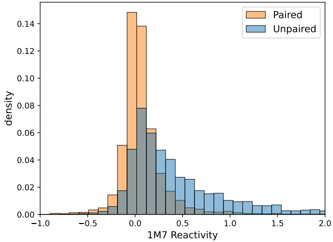
Secondary structure is not uniquely determined by reactivity, even with perfect data!

Reactivity Distributions

How well do reactivities distinguish paired from unpaired?

Reactivity Distributions

How well do reactivities distinguish paired from unpaired?



- Paired positions less reactive than unpaired
- Distributions overlap strongly
- Best distinction only for very large reactivities
- Negative values due to noise

Structure Prediction with Reactivities

How can we incorporate reactivities into structure prediction?

- Sample & Select:
  - Predict candidate structures
  - Select candidate that best fits measured reactivities
    E.g. candidate s that maximizes $P(r|s)$
- Soft constraints:
  - Derive a pseudo-energy from reactivities
  - Modifies the energy model to prefer structures that fit the data

Deigan's Pseudo Energies

Position-dependent pseudo energy applied to stacking energies:

$$\Delta G = m \ln[1 + r_i] + b$$

- first implemented in `RNAstructure`
- most widely used method
- works (surprisingly) well, but ...
  1. m and b have to fitted by probing known structures
  2. Why change energy for positions that are already predicted correctly?
  3. there is no good interpretation of the folding energies with the pseudo energies

Zarringhalam Pseudo Energy

The Deigan pseudo-energy can make prediction worse by driving the structure ensemble *away* from measured data.

1. Suppose position i is predicted to be 80% unpaired
2. Shape data suggest i is only 70% unpaired
3. Pseudo-energy push towards even higher unpaired probability

Zarringhalam & Clote suggest adding a pseudo energy for structure π, given probability q_i (from measurement) that pos i is unpaired:

$$\Delta G_k(\pi) = \sum_{i=1}^{n} \beta \, |\pi_i - q_i|, \quad \pi_i = 1 \text{ if } i \text{ unpaired, } 0 \text{ if paired}$$

This is *guaranteed* to always bring the ensemble closer to the measurement.
Requires a model to convert reactivities r_i into a probability to be unpaired q_i

Washietl Approach

- Both $\vec{q}$ (measurement) and NN energies are have errors.
- Predict the probability p_i that pos i remains unpaired. From this compute the discrepancy between measurement and prediction $\|\vec{p} - \vec{q}\|$.
- **Task:** Find a pseudo-energy that is (i) small and (ii) minimizes the discrepancy between prediction and data. Compute energy correction ϵ_μ that minimize

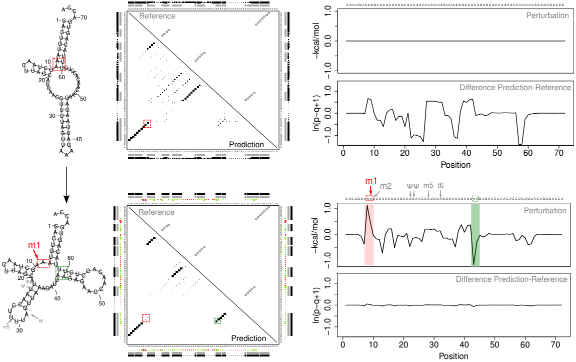
$$F(\vec{\epsilon}) = \sum_{\mu} \frac{\epsilon_\mu^2}{\tau_\mu^2} + \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left(p_i(\vec{\epsilon}) - q_i \right)^2$$

σ and τ encode our trust in the energy parameters and experimental data.
Ensemble based – does not assume a single structure

Making use of the Perturbation Vector

Example: tRNA modifications

- Human mitochondrial tRNA-Lys does not fold correctly *in vitro*
- Methylation at position 9 restores folding to the cloverleaf shape



Probabilistic Approach

Let $P(r|\pi)$ be the likelihood of observing the reactivity vector r given structure π on sequence x. The posterior probability of structure π is

$$P(\pi|r, x) = \frac{P(r|\pi, x) \cdot P(\pi_i|x)}{p(r)}.$$

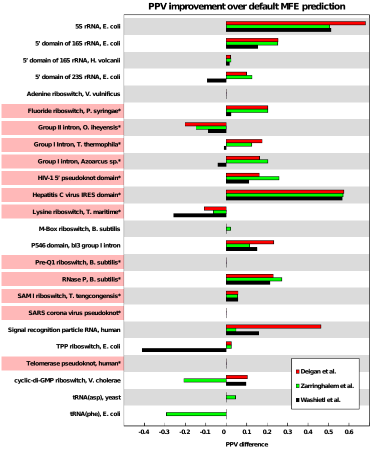
where the prior $P(\pi|x)$ is the Boltzmann probability of π
Assuming that r_i only depends on the structure state π_i at pos i, maximizing $P(\pi|r, x)$ is equivalent to adding a pseudo energy

$$\Delta G(\pi_i, i) = -RT \log P(r_i|\pi_i)$$

- First proposed by Sean Eddy
- We're free choose which structure states to consider
e.g. three state model $\pi_i \in \{\mathrm{unpaired}, \mathrm{stacked}, \mathrm{helix\ end}\}$
- Still assumes that only a single structure π is present

Probing Data in ViennaRNA

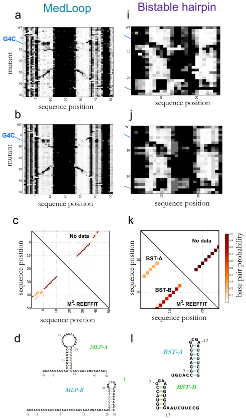
- All four pseudo-energy methods supported in ViennaRNA
- Probing data don’t always improve prediction



Mutate-and-Map

Probing looks at single positions, how to learn about base pairs?

- Mutate every position
- When mutation at i breaks a pair (i, j), reactivity of the partner j changes!
- $\rightarrow$ direct information about pairs (i, j)
- sometimes mutations cause complete refolding
- can identify alternative structures



Cordero, . . . , Das (2014)

Structure Ensembles

What if our RNA can form multiple structures?

- Measured reactivity is an average over ensemble

$$r = \sum_{\pi} p(\pi) \cdot r(\pi)$$

Possible solutions:

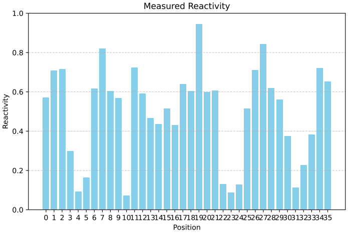
- With a set of candidate structures: estimate $p(\pi)$, minimize discrepancy between observed and expected reactivity
- Need to guess candidates correctly

- Separate reads into clusters corresponding to different structures
- Assumption: Each read is produced by one structure in the ensemble
  All mutations in a read derive from the same structure

Structure Ensembles

Two structure example:

```
((((((....))))))....((((((....))))))     ...((((((((.....(((...))).))))))))..
GAAAGCC-G-GCUUUU--C-CCG-C-CAU-GGCUGG    GAAAGCCUGU-CUUUUGCCA--GG-UCAUGGGCU-G
GAAAGC-UG-GCUUUUG-C-CCGGCU-AUGGGCUGG    G-AAG-CUGUGCUUUUGCC-CCGG-U-AUGGGCU-G
GAAAGC-UGUGCUUUUGCCACCGGCUCAUG-GCUGG    ---AGCCUGUGCUUUUGCCAC-GGCUCAUGGGCUGG
GAAAGC--GUGCUUUUG--ACCGGCUCA-GGGCUGG    GA-AGCCUGUG-UUUUGCCACCGGC-CAUGGGCUG-
GAAAGC-U-UGCUUUUGCCACCGGCUCA--G-CUGG    GAAAGCCU-U-C-UU-GCCA--GGCUCAUGGGCUGG
GAAAGCCUGUGCUUUU-CC-CCGGCUC--GGGCUGG    -AAAGCCUGUGCUUUU-CCACCGGC-CAUGGGCU-G
GAAAGCCUGUGC-UUUGCCACCGGCUC--G-GCUGG    GA-AGCCUGUGC-U-UG-CA-CGGCU-AUGGG--GG
GAAAGC---UGCUUUUG-CACC-GCUCAUGGGCUGG    GA-AGCCUGUGCUUUUG-CACCGGCUCAUG-GCU-G
GAA-GCCUGUGCUUU----ACCG-CUC-UGGGCUGG    GAAAGCCUGUGCUUU-GCCACCGGCUCAUGGGCUGG
GA--GCC-GUGCUUUUG-C-CCGGCUCAUGGGCUGG    -AAAGCCUGUGCUUUUGCCAC-GGCU-A-G-G-U-G
```

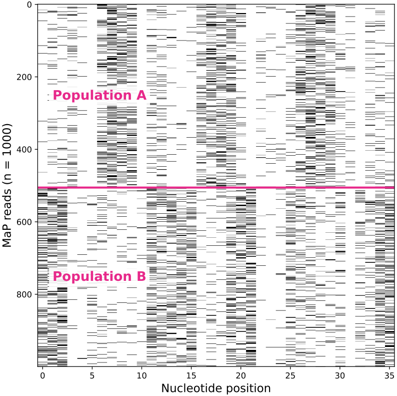


Structure Ensembles

Two structure example:

```
((((((....))))))....((((((....))))))
GAAAGCC-G-GCUUUU--C-CCG-C-CAU-GGCUGG
GAAAGC-UG-GCUUUUG-C-CCGGCU-AUGGGCUGG
GAAAGC-UGUGCUUUUGCCACCGGCUCAUG-GCUGG
GAAAGC--GUGCUUUUG--ACCGGCUCA-GGGCUGG
GAAAGC-U-UGCUUUUGCCACCGGCUCA--G-CUGG
GAAAGCCUGUGCUUUU-CC-CCGGCUC--GGGCUGG
GAAAGCCUGUGC-UUUGCCACCGGCUC--G-GCUGG
GAAAGC---UGCUUUUG-CACC-GCUCAUGGGCUGG
GAA-GCCUGUGCUUU----ACCG-CUC-UGGGCUGG
GA--GCC-GUGCUUUUG-C-CCGGCUCAUGGGCUGG
```

```
...((((((((.....(((...))).))))))))..
GAAAGCCUGU-CUUUUGCCA--GG-UCAUGGGCU-G
G-AAG-CUGUGCUUUUGCC-CCGG-U-AUGGGCU-G
---AGCCUGUGCUUUUGCCAC-GGCUCAUGGGCUGG
GA-AGCCUGUG-UUUUGCCACCGGC-CAUGGGCUG-
GAAAGCCU-U-C-UU-GCCA--GGCUCAUGGGCUGG
-AAAGCCUGUGCUUUU-CCACCGGC-CAUGGGCU-G
GA-AGCCUGUGC-U-UG-CA-CGGCU-AUGGG--GG
GA-AGCCUGUGCUUUUG-CACCGGCUCAUG-GCU-G
GAAAGCCUGUGCUUU-GCCACCGGCUCAUGGGCUGG
-AAAGCCUGUGCUUUUGCCAC-GGCU-A-G-G-U-G
```



- Reads from the same structure exhibit similar mutation patterns
- Reads can be separated into clusters belonging to different structures

Methods for Ensemble Deconvolution

Several methods exist to deconvolute reads

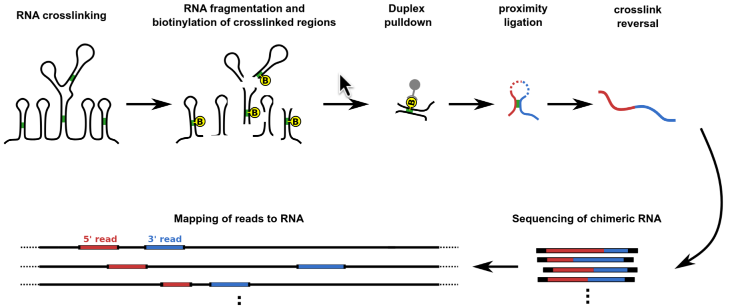
Method	Clustering algorithm	Number of clusters
DREEM	Expectation-Maximization	User specified
DRACO	Spectral Clustering	Automatic
DANCE-MaP	Expectation-Maximization	Automatic

After deconvolution the structure corresponding to each cluster is predicted separately

In the ideal case long-range tertiary interactions can be detected by correlation between sites

RNA Crosslinking

RNA crosslinking can directly detect RNA-RNA interactions



- Psoralen induced crosslinking and pulldown
- → chimeric reads corresponding to two interacting regions
- Should give information on (hard to predict) long-range base pairs

RNA Crosslinking

- Ideal for detecting RNA-RNA interaction (read-pair from two different RNAs)
- Challenges:
  - No nucleotide resolution:
    Interaction could be anywhere between the regions
  - Cross-linking only implies that regions are closed, not base paired
  - High noise:
    Many reads do not correspond to true interactions
    In read-pairs from human 18S/28S rRNA, more than 50% false positives

Not widely used yet for secondary structure determination
DRACO and other programs can combine cross-linking and SHAPE-Map data in a single analysis

Take Home Messages

- Chemical probing is a fast and inexpensive
- Can significantly improve structure prediction
- Probing data are noisy and differ in quality
- Probing signal is affected by other factors
    - accessibility of a site in 3D structure
    - non-canonical base pairs
    - tertiary interactions
- Structure ensembles complicate analysis
  More reads and more mutations per read needed for deconvolution